White Paper Report

Report ID: 102649

Application Number: HD-51269-11

Project Director: Will Hanley (whanley@fsu.edu)

Institution: Florida State University

Reporting Period: 5/1/2011-8/31/2014

Report Due: 11/30/2014

Date Submitted: 11/30/2014

Final Performance Report
Grant Number: HD-51269-11
Project Title: Populating Prosop, A Social Networking Tool for the Past: Two Workshops
Project Director: Will Hanley, Florida State University
November 30, 2014

Narrative Description

Project Activities

We undertook three major activities during the grant period: prototype development and two workshops.

1) Prototype development was the most challenging aspect of the project. Prosop began as an idea for a highly flexible database of historical individuals. Early in development, it became clear that this project could best be accomplished by using a database structure that had not yet been adapted for use in a digital historical names project. That data structure is the Resource Description Framework (RDF) model. RDF has several distinct advantages over the relational and xml data models used by existing historical names projects: it does not require a pre-determined data structure and works well with fragmentary and ambiguous evidence. RDF also has a distinct disadvantage: thus far, it has not been used in any such application. As the project progressed, it became clear that simply producing a presentable, functioning example of an RDF database of historical person data would be a significant contribution to historians seeking such a tool.[1]

In the early stages of the grant, we worked to find technical collaborators in computer science who would be willing to design a RDF database for us. An NSF grant proposal in which Prosop was a "driving project" went unfunded, and we began to search for other development avenues. The cost of hiring a postdoc or graduate student in computer science was well beyond our budget. Although several people were interested in applying for new grants together, it proved difficult to find someone at a university who was willing to do the technical work on the existing grant.

Fortunately, a graduate student in the Florida State University history department named Chris Osmar offered to use his programming background to contribute to the project, and the department contributed two research assistantships to support his work. He worked for many months, in fits and starts as time allowed, to produce a workable prototype. He used C++ to write a new program that became a functional import wizard, able to convert csv spreadsheets into semantic web data classified according to user-defined categories and labeled by source. He developed a "factoid" model of data that proved to be a useful concept moving forward. We hope to publish a brief paper describing the features of this undertaking.

Although we held the first workshop (described below) on the basis of this preliminary prototype, ultimately there were too many pieces to build for us to continue to build everything ourselves. At the same time, we remained convinced that the project would only have an impact if we managed to present a complete RDF database. Also, workshop participants and other colleagues insisted on the wisdom of using externally supported software. We therefore began another search for an existing product or project that could meet our needs. This time, we found stardog (stardog.com), a commercial graph database. We made contact with Clark & Parsia, the company that produces the database, and found that Kendall Clark has a background and interest in the humanities and was willing to work with us.

During the first six months of 2014, we engaged in many conversations with Kendall Clark and other Clark & Parsia staff. Kendall did an excellent job of translating our humanities concerns into the terms

1 As one critic of RDF put it, "The Semantic Web was a *great* idea in 2003. The Semantic Web is still a great *idea* in 2013." Robert Sanderson, "RDF: Resource Description Failures and Linked Data Letdowns," *Journal of Digital Humanities* 2.3 (Summer 2013).

of database design, and offered tremendous help in defining our needs and expectations and helping us to understand our limitations. At present, we have a basic prototype which works on some of the archival data produced by Will Hanley concerning Algerian French subjects registered at the French consulate in Algeria between 1873 and 1914. We use OpenRefine as an import wizard, cleaning and formatting the data as we bring it into RDF. We have been able to produce some basic visualization and to define a basic ontology of name rules based on this data. Our work together is ongoing, and we hope to make this material public. We also have some hope of collaborating on grant applications with Clark & Parsia in the future. The prototype does not yet make a persuasive case for the use of the graph data model in the digital humanities, but it should be so in the near future.

2) The grant proposal envisaged two workshops that would refine and populate a Prosop prototype. Development of the prototype was far more difficult than we had imagined. After receiving two extensions, and still without a working prototype, we decided to proceed with the first Prosop workshop, which was held on May 17-18, 2013 at Brown University. The call for participants yielded twenty-five applications, ten of whom we invited to participate. A dozen other members of the Brown community and the broader public also joined in the conference. The conference was hosted and supported by Middle East Studies at Brown. Thanks to Beshara Doumani (director), Tony Watson (assistant director), Elias Muhanna, and especially Barbara Oberkoetter for their generous hospitality.

The workshop program comprised presentations from eleven participants, as well as joint discussions of the proposed structure of the tool. Participants pre-circulated tranches of their own databases, showing about 100 person-records, and fitting these tranches into Prosop was the focus of presentations and discussions. This report is a brief summary of some of the salient points that arose.

After a presentation of the database scheme, the first session featured three archivists and editors, who discussed their databases of New Englanders.

- Joanne Riley (Umass Boston) introduced us to a database of life insurance policies issued to 26,789 members of the Catholic Order of Foresters between 1879 and 1935. Names and dates of this collection were keyed in by volunteers of The Irish Ancestoral Research Association, and it has been digitized by familysearch.org. The fuzziest data from the life insurance policies, its medical details, has not been entered into the database. Many of these questions (for example, “Do you ride a bicycle?”) could be of interest to social historians.
- Nicole Topich (Center for American Political Studies, Harvard) is project archivist of the Digital Archive of Massachusetts Anti-Slavery and Anti-Segregation Petitions, which was awarded an NEH Humanities Collections and Reference Resources Foundation grant in March 2012. This project is digitizing and making a database of 4-6,000 petitions, many of which contain tens of thousands of signatures. This material will be available online, and the project aims to include many additional tags (such as geodata).
- Ondine LeBlanc (Massachusetts Historical Society) presented the Colonial Collegians project, which is based on a collection of narrative biographies of 6,000 attendees of colonial colleges between 1642 and 1774. A name/date directory already exists, but the MHS wants to make the sketches searchable in their full richness, in order to reveal (for example) the religious history and the female networks latent in the text.

The second session featured two well-established person databases.

- Jean-Pierre Dedieu (Lyon) spoke about Actoz, a database of 100,000 Spanish political actors from the 17th, 18th, and 19th centuries. This is an application of a Filemaker based program that he has developed.

- David Radcliffe (Virginia Tech) Lord Byron and his Times, 10,400 names from Lord Byron's correspondence, early 19th century

Friday's third session included two speakers whose databases center around text objects that contain names.

- John Nielson (Benedictine) Names from 700 Neo-Babylonian legal and administrative tablets, 747-626 BC
- Dagmar Riedel (Columbia) Encyclopedia Iranica and manuscript cataloging

On Saturday, we had four presentations:

- Heather Lee (American Studies, Brown) presented 950 Chinese-American draft cards from World Wars One and Two that form part of her ongoing research on Chinese restaurants in North America. She has found it impossible to adequately disambiguate the mass of unsystematically-transliterated names that she finds in the archive. As a result, she has organized her material with date and point of collection as the primary category. She made the provocative suggestion that this document-focused archival epistemology (which accords with other important research) could be generalized, for instance in Prosop. The implications of this suggestion are difficult to grasp at this point.
- Elaine Parsons (Duquesne) 5,705 names from Union County, South Carolina criminal indictments, 1852-1878
- Karen Wilson (UCLA) 2,201 members of California Jewish networks, c1850-1900
- Will Hanley (Florida State) 10,000 names from Egyptian legal records, 1880-1914

Our most extended discussions concerned the question of confidence score. Some noted that more knowledgeable users might assess their data as less reliable. Other comments about Prosop suggested that it must better acknowledge the financial costs of long-term, evolving data storage. Also, it was suggested that any such system should be based on a widely-used, commercial format in order to assure future convertibility. This is particularly true if Prosop is to fulfill its preservationist aspirations.

A few more conclusions:
- there are many risks associated with acting as a primary data receptacle. It may make sense to suggest that Prosop can (also) act as a secondary or parallel database, for the names component of data projects with other aims. It might be difficult, for example, for Prosop to convey the archival context of every kind of person record.
- unstructured data can survive obsolescence better than fixed programs. Prosop aims to preserve barely structured data, while providing a superstructure for its experimental organization.
- an export wizard is necessary.
- every project uses unique identifier numbers for persons. Can semantic web URIs, which might not so clearly emphasize the primacy of the person, organize the data as well?

Overall, most of the participants strongly endorsed the need for a semantic web tool for historical name data. There was considerable wariness about the obstacles to realizing such an objective, however: if it made so much sense, why hadn't someone done it yet? Armed with insights from the vivid discussions, we refocused our aim on adapting an available tool for our purposes.

3) The second workshop, hosted at Florida State University, was not as successful as the first. An initial call for participants for a meeting on May 9, 2014 yielded only two applications. This date was quite close in the calendar year to the previous workshop; perhaps weaker publicity, the decreased novelty of

the project, or the more remote location accounted for the smaller return. We decided not to hold a workshop in the spring, preferring to hold a workshop at the end of the summer. On August 15, 2014, we held this workshop.

A dozen people attended the workshop, which featured an introduction to the RDF data structure from Héctor Pérez Urbina (Clark & Parsia) and a discussion of the Prosop prototype (Will Hanley, Héctor Pérez Urbina, Chris Osmar). We then looked at three use cases: Alexandria data (Will Hanley), the China Biographical Database Project (Michael Fuller, UC Irvine), and the Social Geography of the Islamic World (661-1300 CE) (Maxim Romanov, Tufts).

Participants found the workshop an excellent opportunity to learn about RDF person databases and to exchange experiences with counterpart projects. Unfortunately, the relatively weak response and the relatively simple state of the prototype did not allow us to do the populating work that was envisaged in the grant proposal.

Accomplishments

Through consultation and discussion with a broad variety of historians and computer scientists, through presentations, workshops, and one-on-one meetings, we were able to articulate a project description, refined as follows:

> **I Primary function: person-records**
> Each person-record will consist of literal entries in one or more fields. Each field will belong to a defined category. For each field entry, the database must also record four additional items: author, attribution, confidence score, and permissions. Other users can add additional fields to an existing person-record, without changing the original.
>
> **Category**: Every field will be given a literal name, and related to one or more "registered" categories. The relation and the registered categories themselves must be changeable. The system administrator will establish a canon of registered categories, each of which will be related to a data type and a means of translating literal records into that data type, without changing the original.
>
> For example, an archival record might state that someone "passed away on the first day of June in 1910." The category should be "passed away" and the record should be "the first day of June in 1910." This category could be related to a canonical category "Death Date," with data format DD.MM.YYYY, requiring an entry of 01.06.1910. It could later be related to a second canonical category "Deceased When" with data format MM.DD.YYYY, requiring an entry of 06.01.1910.
>
> A more advanced example of category registration: an archival record might state that someone "worked at the raising of pigs." The literal category would be "worked at" and the literal record would be "raising of pigs." The category could be related to a canonical category "Employment Field," with a fixed list of 100 occupations, from which this record would be assigned to "Animal Husbandry." Here the system could employ existing standards, such as the Historical International Standard for Classification of Occupations (HISCO). It is by this means that linguistic translation will be accomplished.
>
> The database must also tolerate contradictory categories. A field can be associated with multiple possible but mutually exclusive categories. For example, two person-records about "John Baker" and "Sam Barber" could classify "Baker" and "Barber" as both family names and occupations.

Author: the name of the user contributing the record, and the time when it was contributed, will be recorded. All subsequent edits, including by other authors, will be recorded in similar manner. This can resemble the "edit record" of a wiki. Every change will be recorded at the level of the field, but also aggregated to a list of all changes to the person-record.

Attribution: every field must be linked to an archival source. This attribution can be literal, but can also (and more frequently) refer to a specific record within a larger registered archive. For example, the "passed away on the first day of June in 1910" record should be attributed to the registered archive "US National Archives College Park" at literal site "449/23/17."

It should be possible to add a second archival source that attests to the same record, where possible.

Confidence score: this evaluation of the quality of the data will be set by the author. It will have a default value, and will be editable.

Permissions: the author can decide what other users can see which parts of the record, according to a system to be worked out. For example, the literal text "raising of pigs" could be hidden but "Animal Husbandry" visible. Permissions must be changeable.

User Interface: Individuals will log-in to a web-based platform. Data will be entered via wizard, or directly via a user-defined form.

II Secondary task: hypotheses and networking

While preserving all original data in its native form, other users will be able to experiment with hypothetical layers. They can, for instance, see what happens if a category is defined differently. For instance, if "raising pigs" is not "Animal Husbandry" but (in a newly-defined category) "Raising non-Kosher Meat," do we in fact see fewer Jews in the occupation? The user would be able to record and publish this hypothetical layer.

A similar kind of work must be possible in networking persons. Users should be able to suggest that two or more person records are related through bonds of kinship, profession, friendship, place, and so on. Users should also be able to suggest a relationship of identity between two records, i.e. that they represent the same person.

All suggestions of this sort should be recorded in a similar fashion to the original records: author of the hypothesis, confidence score, attribution (here it would be to reasoning or other records, rather than to a specific archival source), and permissions.

III Tertiary task: Study of categories themselves.

As more and more records are added, the categories will change. More will become canonical, some canonical categories might be merged, and it will become clear that, for example, some categories are more present at different times and in different places over the whole of the database. It will be necessary to be able to filter the count of use of these categories according to various filters. Every category will have its own page, which displays various information about how that category is used.

Translation: For users viewing the database in another language, canonical categories would display canonical translations. So, for an example given earlier, "worked at the raising of pigs," someone viewing the record in French will see the relationship with "Emploi" and the occupation "Elevage d'animaux." The literal record, in its original language, will still be visible.

IV Tertiary task: Study of networks
The system should display genealogical tables, geographic networks, and so on. It should be able to do so using plug-ins developed by users, which will access the data to which they've been given permission.

V Tertiary task: Author and Source pages
Each registered author and each registered archival source used for attribution will have its own page, recording all of the entries created by the author or using the source and allowing assessment of various kinds of the author or the source.

VI Tertiary task: Confidence score
An algorithm will generate a confidence score for each person-record, and each field within that record, based on the author-defined confidence score, the author's reputation score (generated using the author's page and record of contributions), the source quality score, and any secondary verification.

Most of these points can be satisfied through the RDF data model, managed by stardog software, using Google Refine as the import wizard. We do not yet have a convincing, standalone prototype, populated with data. When we achieve this, we will be able to populate it with data gathered at workshops and by other historians interested in our project. The great challenge that remains, however, is the design, testing, and refinement of an ontology that will cover a broad spread of human history. If we had the grant to write again, that task would be the core of the workshop proposal, but as is often the case we only discovered our real aim at the end of the work.

Audiences

The principal investigator made numerous formal and informal presentations of this project. Some of the formal presentations included:
- FSU digital scholars group, 2012
- Routes and Roots conference (University of Toronto), 2012 and 2014
- Middle East Digital Humanities conference (Brown University), 2013
- Orient-Institut Beirut, 2014
- Moving People, Linking Lives symposium (University of Virginia), 2015

The project website (prosop.org) saw considerable traffic when it was first established in 2011 and during the leadup to the first workshop in 2013. It should attract more views when we publish some results of our first technical work and when the prototype and its data sees public release.

Evaluation

Prosop did not have an external evaluation. Some sense of internal evaluation appears in the preceding lines. Stated more directly, this project faced a number of external and internal challenges. Externally, the challenge of building an RDF person database was much more complex than originally imagined, and the challenge of finding competent technical collaborators willing to work for a relatively modest sum of money was also greater than imagined. These were challenges of inexperience. The internal challenges had mostly to do with the PI's lack of flair in management and his attempts to do work himself that was beyond his competence. Also, in an institutional sense it proved difficult to prioritize work on a digital humanities project that would receive little credit in tenure evaluation.

Those difficulties aside, it seems clear that Prosop has developed a workable, nuanced RDF model for the management of historical person data that is realizable through stardog and should (when fully formed) support innovative historical work. If this goal has not yet been achieved, its potential should soon be more convincingly demonstrated by a released beta demonstration.

We requested and received three extensions for this grant. The original deadline envisaged accomplishing the project over four summer months. We greatly underestimated the technical challenges that the project involved, as well as the costs of running any digital humanities project with a non-trivial software development component.

Continuation of the Project

We will finish work with Clark & Parsia to release polished beta demonstration app as described in "Final product and dissemination " section of the original grant application: (The final product for this start-up grant will be a populated prototype of Prosop that can be released to the public in a beta version. The tool will have dozens of trained users and tens of thousands of records.) We have a number of contacts willing to let their data be used to populate this prototype, and we will produce some basic visualizations of this material. This is proof of concept necessary to move forward from the start-up phase. Will Hanley has a couple of talks scheduled for 2015, and will use this beta to demonstrate the actual workings of the Prosop tool. We have had several conversations with various parties about seeking further funding. From our perspective, the next priority is ontology design.

Long Term Impact

Prosop was the first project at Florida State University to receive an NEH digital humanities startup grant. There were few facilities to support this work when we received the grant, but that situation is gradually changing, helped (indirectly) by the award of the grant itself, which made it clear that digital humanities work is part of the faculty research agenda at FSU. The university has since received another DH startup grant, and the library has hired a number of digital humanities librarians to support faculty and student research. As the library has moved to provide better services, I've talked with them frequently about digital humanities needs. Thanks to their good work, the university is in a much better place for such research than it was in 2011.

Grant Products

At present, the only publicly available grant product is the material available at prosop.org. That material is due for an update, and it will include visualizations from the prototype.

Will Hanley is working on a thought paper on the potential of large volume databases like Prosop for modern microhistory, called "Digital History from Below."